D946–D950 *Nucleic Acids Research, 2009, Vol. 37, Database issue* *Published online 25 October 2008*
doi:10.1093/nar/gkn819

GOBASE: an organelle genome database

Emmet A. O'Brien*, Yue Zhang, Eric Wang, Veronique Marie, Wole Badejoko, B. Franz Lang and Gertraud Burger

Robert-Cedergren Center for Bioinformatics and Genomics, Département de Biochimie, Pavillon Roger-Gaudry, Université de Montréal, 2900 Edouard-Montpetit, Montreal QC, Canada H3T 1J4

Received September 11, 2008; Revised October 10, 2008; Accepted October 13, 2008

ABSTRACT

The organelle genome database GOBASE, now in its 21st release (June 2008), contains all published mitochondrion-encoded sequences (~913 000) and chloroplast-encoded sequences (~250 000) from a wide range of eukaryotic taxa. For all sequences, information on related genes, exons, introns, gene products and taxonomy is available, as well as selected genome maps and RNA secondary structures. Recent major enhancements to database functionality include: (i) addition of an interface for RNA editing data, with substitutions, insertions and deletions displayed using multiple alignments; (ii) addition of medically relevant information, such as haplotypes, SNPs and associated disease states, to human mitochondrial sequence data; (iii) addition of fully reannotated genome sequences for *Escherichia coli* and *Nostoc* sp., for reference and comparison; and (iv) a number of interface enhancements, such as the availability of both genomic and gene-coding sequence downloads, and a more sophisticated literature reference search functionality with links to PubMed where available. Future projects include the transfer of GOBASE features to NCBI/GenBank, allowing long-term preservation of accumulated expert information. The GOBASE database can be found at http://gobase.bcm.umontreal.ca/. Queries about custom and large-scale data retrievals should be addressed to gobase@bch.umontreal.ca.

INTRODUCTION

The amount of information available in generalist molecular sequence databases such as GenBank (1) continues to grow, and this information becomes more diverse and complex as we discover new biological phenomena. Therefore, there is an increasing need for expert databases specializing in particular areas of molecular biology. Specialist databases provide expert curation of data, and access to that data in a flexible and well-integrated fashion serves a purpose complementary to generalist databases such as GenBank.

GOBASE is one such specialist database, which has been collecting, curating and publishing data concerning mitochondrial and chloroplast genomes since 1995 (2–5). Organelle genomes are of biological interest for a wide range of studies, such as molecular taxonomy, molecular mechanisms of trans-splicing and RNA editing, and non-Mendelian inherited metabolism-related disease in humans. GOBASE contains a number of different categories of data, such as nucleic acid and protein sequences, genetic maps, taxonomic data and RNA secondary structures. All gene and product names have been assigned from a locally maintained standard list, and this combines with a powerful and flexible interface to allow a wide range of complex searches. While initially GOBASE was designed primarily to address issues of comparative biology, such as the diversity of organelle genome structure in eukaryotes (e.g. 6,7), we have more recently added functionality specific to the human mitochondrial genome in GOBASE, such as searches by haplotype and disease state, which are of medical interest.

DATA CONTENT

GOBASE release 21 (June 2008) contains 913 000 mitochondrial sequences including 737 000 genes, and 250 000 chloroplast-encoded sequences including 174 000 genes, derived mostly from GenBank releases up to 164. The large number of complete organelle genomes available makes GOBASE a valuable resource for phylogenomics, with 6300 complete mitochondrial genomes and 213 chloroplast genomes. This number has increased almost 4-fold since the previous report.

More recently (5), we have added bacterial genome sequences for reference purposes. As of release 21 GOBASE includes three complete bacterial genomes: *Escherichia coli* K12; the alpha-proteobacterium

*To whom correspondence should be addressed. Tel: +1 514 343 6111; Fax: +1 514 343 2210; Email: eobrien@bch.umontreal.ca

Rickettsia prowazekii strain Madrid E, closely related to the bacterial ancestor of mitochondria; and the cyanobacterium *Nostoc* sp., closely related to the bacterial ancestor of chloroplasts. In order to provide a consistent comparative view of these genomes, they have each been reannotated using the AutoFACT functional annotation tool (8), including assignation of Gene Ontology terms. GOBASE now contains 10 700 bacterial genes in total.

ENHANCEMENTS TO FUNCTIONALITY

RNA editing

RNA editing refers to a molecular process by which the sequence of a transcribed RNA is modified. This has been seen to occur in the mitochondria of several eukaryotic taxa, such as plants (9) and trypanosomes (10), and in chloroplasts (11). At the level of basic changes, examples exist in the database of sequences being modified by the substitution of one residue for another, by deletion of residues, and by the addition of residues, usually uracil.

The *RNA editing* interface in GOBASE is based primarily on the previously existing *RNA* query page, with the addition of editing-specific selection parameters such as the type of modification (insertion, deletion or substitution). A query result is shown in Figure 1. In addition to the sequence itself, edited positions are displayed, both as a list specifying the exact change made at each position, and marked in red on an alignment of the relevant sections of sequence for a straightforward and intuitive visual representation. The interface displays only the regions of the sequence where editing occurs. Coding and intronic regions of the sequence are distinguished by background color. Complete unedited and edited sequences can be downloaded from the interface page. Future development will include the possibility of downloading the sequence alignment as displayed, and the addition of multiple rows

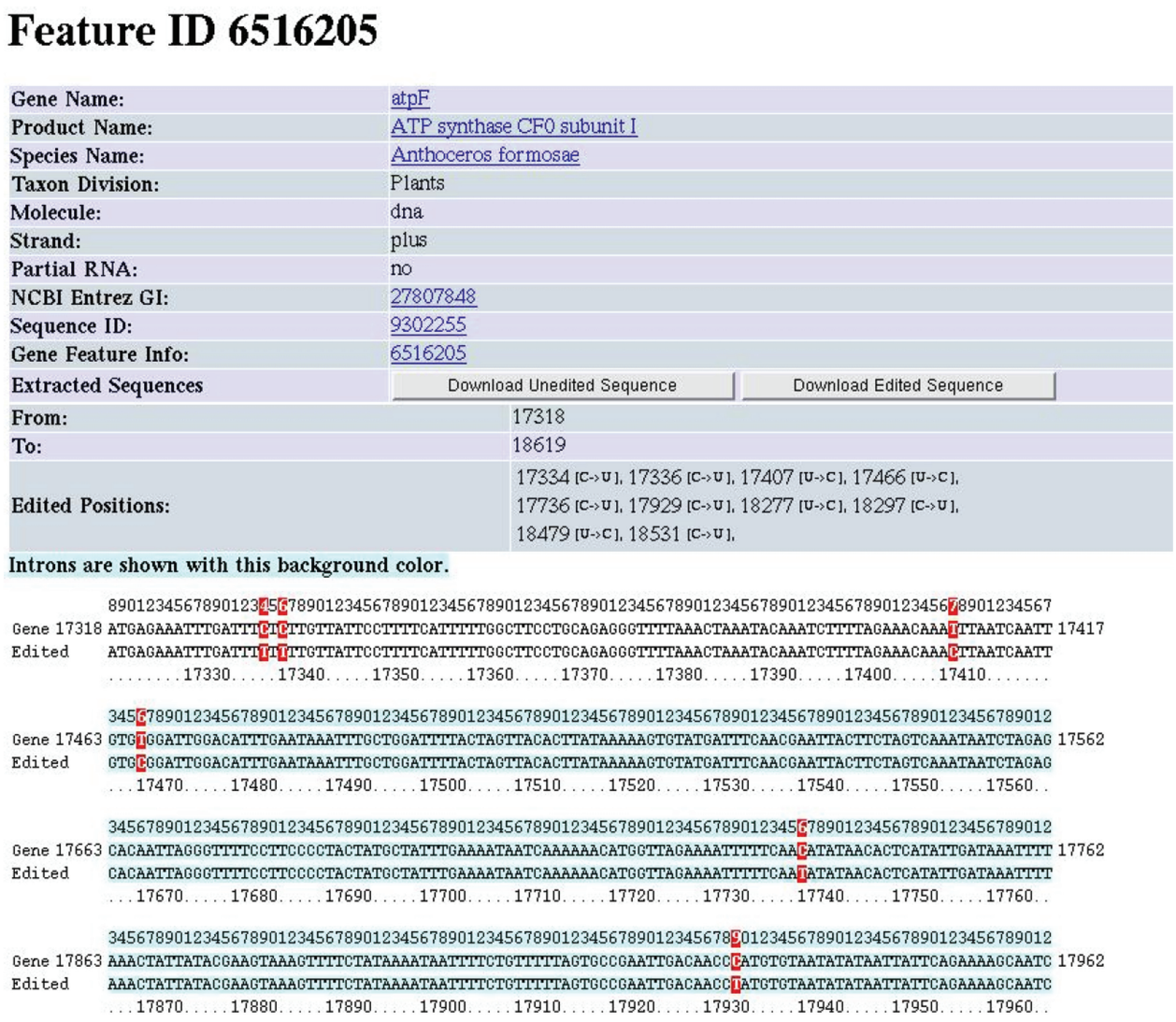


Figure 1. *RNA editing* result page, showing sequence-specific data, location of edited positions and alignment of gene sequence with edited sequence. Hyperlinks lead to database pages for details of appropriate *Gene Product*, *Taxonomy*, *Sequence* and *Gene*, and to the Entrez page for the appropriate gi. Start and end positions of the gene, and locations of edited positions, are numbered relative to the start of the sequence entry containing the gene.

to the alignment in cases where edits to a sequence are known to occur sequentially, so that observed intermediate stages in the editing process can be represented.

Human-specific data

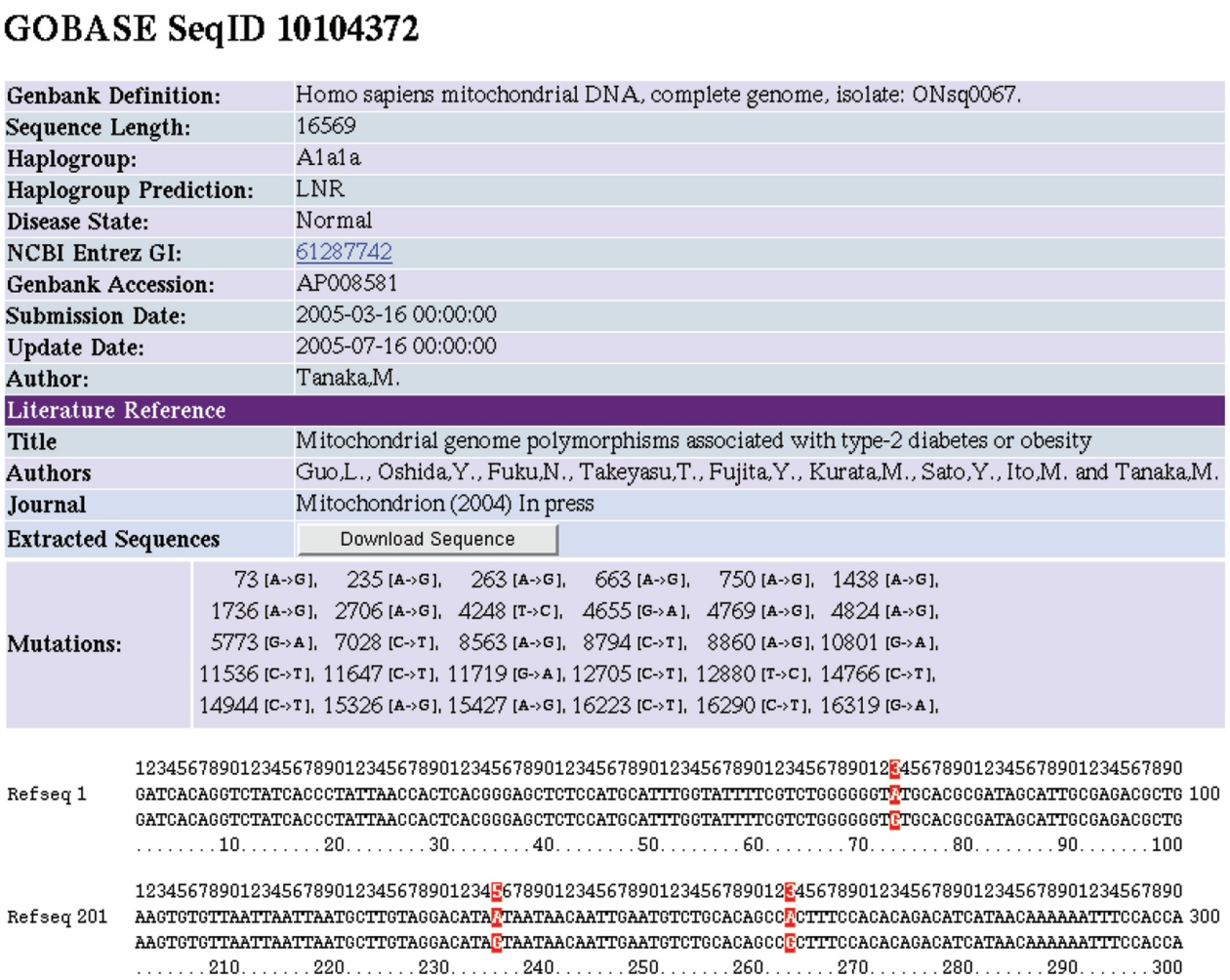
Information specific to the ~3000 complete human mitochondrial genome sequences in GOBASE has been added from a number of sources, including HmtDB (http://www.hmtdb/uniba.it/) (12), OMIM (http://www.ncbi.nlm.nih.gov/omim/) (13) and MitoMap (http://www.mitomap.org/) (14). Two different interface pages provide access to these new data.

The *Human Sequence* query page allows the user to select a set of human mitochondrial sequences based on haplogroup and disease state. More than 450 different haplogroup assignments are available in GOBASE, so a full list might become unwieldy for some queries. As haplogroup designators always start with a letter, the user is offered the option of first selecting an initial letter or letters, and then picking a range of individual haplogroups from the corresponding subset of haplogroup assignments shown in a menu. The results page (Figure 2) provides relevant information from the standard GOBASE *Sequence* page, and also shows all the positions at which this sequence differs from the reference human mitochondrial genome as defined in GenBank (accession no NC_001807) using an alignment. On this alignment, mutations that have been associated with disease are marked in yellow, and other polymorphic mutations are indicated in red.

The *Human Mutation* query page (Figure 3a) allows the user to search the dataset for mutations of interest within a specified range of positions on the human mitochondrial genome sequence, either by specifying start and end positions directly or by selecting one or more genes from a list on the interface. This search returns a list of positions at which mutations are documented. For each mutation (Figure 3b), the result page provides data on its disease associations, a section of the reference sequence showing the location and neighborhood of the mutation, and a list of the sequences in GOBASE containing this mutation.

Other functional enhancements

The DNA sequence download functionality has been modified to allow the user to download either genomic



GOBASE SeqID 10104372

Genbank Definition:	Homo sapiens mitochondrial DNA, complete genome, isolate: ONsq0067.
Sequence Length:	16569
Haplogroup:	A1a1a
Haplogroup Prediction:	LNR
Disease State:	Normal
NCBI Entrez GI:	61287742
Genbank Accession:	AP008581
Submission Date:	2005-03-16 00:00:00
Update Date:	2005-07-16 00:00:00
Author:	Tanaka,M.
Literature Reference	
Title	Mitochondrial genome polymorphisms associated with type-2 diabetes or obesity
Authors	Guo,L., Oshida,Y., Fuku,N., Takeyasu,T., Fujita,Y., Kurata,M., Sato,Y., Ito,M. and Tanaka,M.
Journal	Mitochondrion (2004) In press
Extracted Sequences	Download Sequence
Mutations:	73 [A->G], 235 [A->G], 263 [A->G], 663 [A->G], 750 [A->G], 1438 [A->G], 1736 [A->G], 2706 [A->G], 4248 [T->C], 4655 [G->A], 4769 [A->G], 4824 [A->G], 5773 [G->A], 7028 [C->T], 8563 [A->G], 8794 [C->T], 8860 [A->G], 10801 [G->A], 11536 [C->T], 11647 [C->T], 11719 [G->A], 12705 [C->T], 12880 [T->C], 14766 [C->T], 14944 [C->T], 15326 [A->G], 15427 [A->G], 16223 [C->T], 16290 [C->T], 16319 [G->A],

```
                1234567890123456789012345678901234567890123456789012345678901234567890123456789012345678901234567890
Refseq 1        GATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCATTTGGTATTTTCGTCTGGGGGGTATGCACGCGATAGCATTGCGAGACGCTG 100
                GATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCATTTGGTATTTTCGTCTGGGGGGTGTGCACGCGATAGCATTGCGAGACGCTG
                .........10.........20.........30.........40.........50.........60.........70.........80.........90.......100

                1234567890123456789012345678901234567890123456789012345678901234567890123456789012345678901234567890
Refseq 201      AAGTGTGTTAATTAATTAATGCTTGTAGGACATAATAATAACAATTGAATGTCTGCACAGCCACTTTCCACACAGACATCATAACAAAAAATTTCCACCA 300
                AAGTGTGTTAATTAATTAATGCTTGTAGGACATAGTAATAACAATTGAATGTCTGCACAGCCGCTTTCCACACAGACATCATAACAAAAAATTTCCACCA
                .......210.......220.......230.......240.......250.......260.......270.......280.......290.......300
```

Figure 2. *Human sequence* result page, showing the difference between the queried sequence and the reference human mitochondrial genome sequence, both as a list of divergent positions and as an alignment of relevant sections of the sequences.

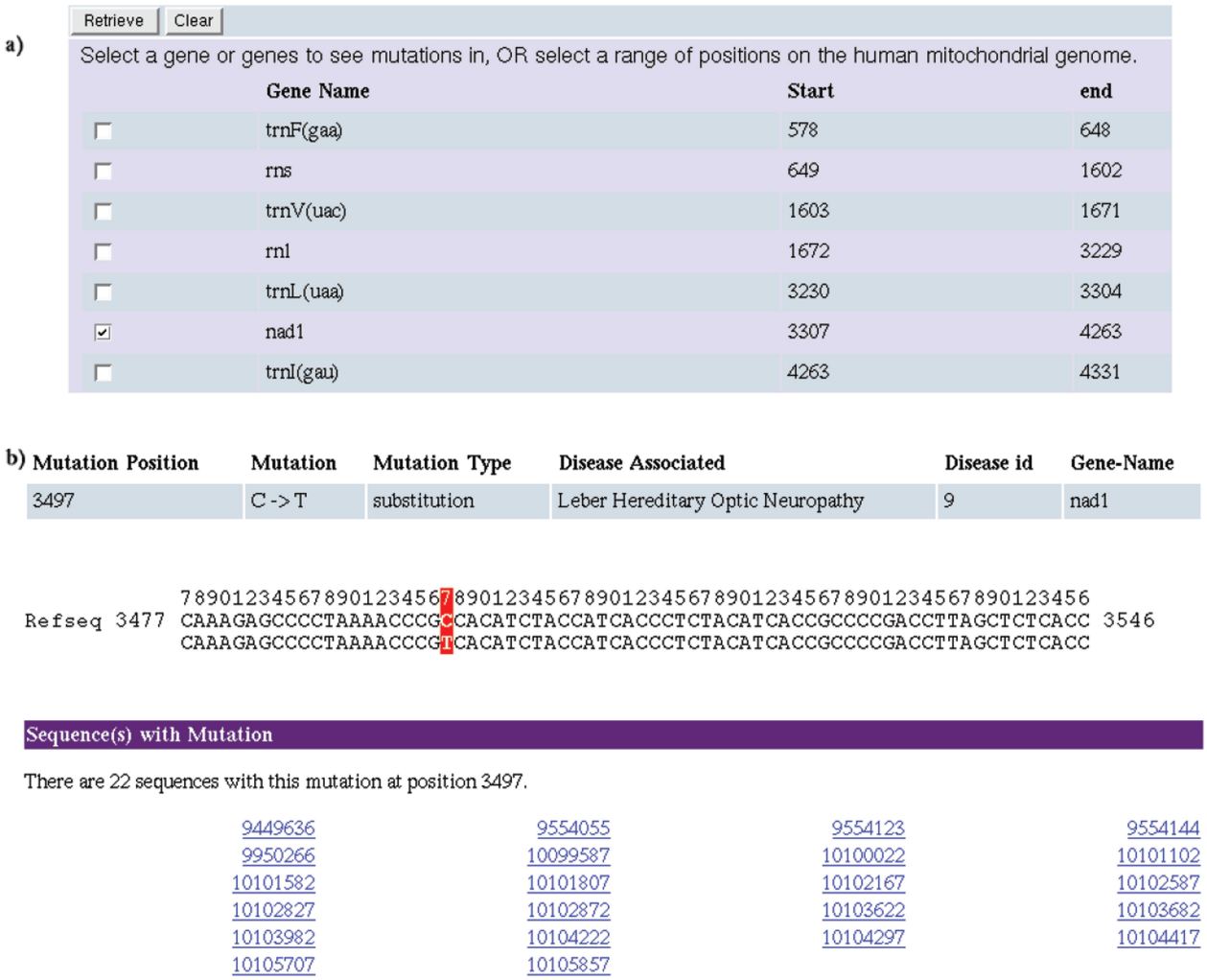


Figure 3. (**a**) *Human mutation* query page, allowing the user to select the gene(s) of interest and specify the range of positions on the sequence to search for mutations. (**b**) Result page showing details for an individual mutation.

sequence or gene-coding regions, selectable via buttons from the *Gene* query page. There are a small number of unusual cases, such as trans-spliced genes, where there is no straightforward correspondence between a single gene and a contiguous linear region of the source sequence record. The GOBASE database structure has now been modified to address these cases transparently. Sequences of complex gene-coding regions are assembled in advance, stored and made available in query results through the same interface as conventional linear genes.

All sequences retrieved from GOBASE now come with detailed literature references derived from the source GenBank records. Journal, author and title are provided, and a direct link to the appropriate PubMed entry if one exists.

Because of practical constraints, any given query in GOBASE returns at most 5000 results. Users wishing to execute custom queries retrieving larger amounts of data are invited to contact the GOBASE team at gobase@bch.umontreal.ca so that the query can be run directly on the database via SQL.

IMPLEMENTATION

The GOBASE database is implemented in version 7.4.1 of the PostgreSQL relational database management system with a web interface written in v4.3.8 of the PHP scripting language. The graphics on the gene pages are generated using the GD module for Perl/PHP, version 2.0.25. Perl (5.8.0) scripts are used to download data from GenBank and process it into GOBASE. All procedures are executed on PCs with two 2.4 GHz or 2.8 GHz Intel Xeon CPUs.

FUTURE PLANS

Specialized databases with all their valuable information are prone to disappearance (15), mostly because of funding constraints, unless transferred to sustainable public databases. We are therefore collaborating with scientists at NCBI to establish a database based on the content of GOBASE as an auxiliary to GenBank. This database will focus on the additional data that expert curation at GOBASE has generated, notably the curated gene and

product names and synonyms and RNA secondary structure data, thus providing a permanent repository for two decades of curation of organelle genome data.

ACKNOWLEDGEMENTS

The authors would like to thank Ilene Mizrachi, Susan Schaefer, Tatiana Tatusova and Jim Ostell at NCBI; Chris Cesaire, Ousman Diallo, and Olivier Tremblay-Savard for contributions to the development of the RNA editing functionality in GOBASE, and Allan Sun for systems administration.

FUNDING

This project was funded by grants MOP-15331 and MOP-84453 from the Canadian Institute for Health Research (CIHR, Genetics Institute). Funding for open access charge: CIHR.

Conflict of interest statement. None declared.

REFERENCES

1. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2008) GenBank. *Nucleic Acids Res.*, **36**, D25–D30.
2. Korab-Laskowska,M., Rioux,P., Brossard,N., Littlejohn,T.G., Gray,M.W., Lang,B.F. and Burger,G. (1998) The Organelle Genome Database Project (GOBASE). *Nucleic Acids Res.*, **26**, 138–144.
3. Shimko,N., Liu,L., Lang,B.F. and Burger,G. (2001) GOBASE: the organelle genome database. *Nucleic Acids Res.*, **29**, 128–132.
4. O'Brien,E.A., Badidi,E., Barbasiewicz,A., deSousa,C., Lang,B.F. and Burger,G. (2003) GOBASE – a database of mitochondrial and chloroplast information. *Nucleic Acids Res.*, **31**, 176–178.
5. O'Brien,E.A., Zhang,Y., Yang,L., Wang,E., Marie,V., Lang,B.F. and Burger,G. (2006) GOBASE – a database of organelle and bacterial genome information. *Nucleic Acids Res.*, **34**, D697–D699.
6. Lang,B.F., Gray,M.W. and Burger,G. (1999) Mitochondrial genome evolution and the orgin of eukaryotes. *Annu. Rev. Genetics.*, **33**, 351–397.
7. Burger,G., Gray,M.W. and Lang,B.F. (2003) Mitochondrial genomes: anything goes. *Trends Genet.*, **19**, 709–716.
8. Koski,L.B., Gray,M.W., Lang,B.F. and Burger,G. (2005) AutoFACT: an automatic functional annotation and classification tool. *BMC Bioinform.*, **6**, 151.
9. Covello,P.S. and Gray,M.W. (1989) RNA editing in plant mitochondria. *Nature*, **341**, 662–666.
10. Benne,R., Van den Burg,J., Brakenhoff,J.P., Sloof,P., Van Boom,J.H. and Tromp,M.C. (1986) Major transcript of the frameshifted coxII gene from trypanosome mitochondria contains four nucleotides that are not encoded in the DNA. *Cell*, **46**, 819–826.
11. Hoch,B., Maier,R.M., Appel,K., Igloi,G.L. and Kössel,H. (1991) Editing of a chloroplast mRNA by creation of an initiation codon. *Nature*, **353**, 178–180.
12. Attimonelli,M., Acceturro,M., Santamaria,M., Lascaro,D., Scioscia,G., Pappad,G., Russo,L., Zanchetta,L. and Tommaseo-Ponzetta,M. (2005) HmtDB, a human mitochondrial genomic resource based on variability studies supporting population genetics and biomedical research. *BMC Bioinform.*, **1**, S4.
13. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., Dicuccio,M., Edgar,R., Federhen,S. *et al*. (2008) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **36**, D13–D21.
14. Ruiz-Pesini,E., Lott,M.T., Procaccio,V., Poole,J.C., Brandon,M.C., Mishmar,D., Yi,C., Kreuziger,J., Baldi,P. and Wallace,D.C. (2007) An enhanced MITOMAP with a global mtDNA mutational phylogeny. *Nucleic Acids Res.*, **35**, D823–D828.
15. Merali,Z. and Giles,G. (2005) Databases in peril. *Nature*, **23**, 1010–1011.